Croatian Operational Research Review
CRORR **7**(2016), 367–388

Decision trees for predicting the academic success of students

Josip Mesarić[1,†] and Dario Šebalj[1]

[1]*Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000 Osijek, Croatia*
E-mail: ⟨{*mesaric, dsebalj*}@*efos.hr*⟩

Abstract. The aim of this paper is to create a model that successfully classifies students into one of two categories, depending on their success at the end of their first academic year, and finding meaningful variables affecting their success. This model is based on information regarding student success in high school and their courses after completing their first year of study, as well as the rank of preferences assigned to the observed faculty, and attempts to classify students into one of the two categories in line with their academic success. Creating a model required collecting data on all undergraduate students enrolled into their second year at the Faculty of Economics, University of Osijek, as well as data on completion of the state exam. These two datasets were combined and used for the model. Several classification algorithms for constructing decision trees were compared and the statistical significance (t-test) of the results was analyzed. Finally, the algorithm that produced the highest accuracy was chosen as the most successful algorithm for modeling the academic success of students. The highest classification rate of 79% was produced using the REPTree decision tree algorithm, but the tree was not as successful in classifying both classes. Therefore, the average rate of classification was calculated for two models that gave the highest total rate of classification, where a higher percentage is achieved using the model relying on the algorithm J48. The most significant variables were total points in the state exam, points from high school and points in the Croatian language exam.

Keywords: Decision trees, academic success of students, classification algorithms, academic performance

Received: September 29, 2016; accepted: December 17, 2016; available online: December 30, 2016

DOI: 10.17535/crorr.2016.0025

1. Introduction

All higher education institutions strive to win over students who are motivated to study and who have a track record in study success. This is usually the

† Corresponding author

presumption for future success. The students enrolled at the Faculty of Economics in Osijek come from various high schools with different education profiles and have also had different levels of success measured according to average grades at schools or state examinations. This, together with their current engagement, probably affects their success in the early phase of their studies. Predicting the success of students in the early phase of their studies helps faculties in directing more activities to less performing students so as to improve their success. According to Simeunović and Preradović [25], analyzing academic success is important for higher education institutions, given that the strategic planning of study programs implies expanding or reducing the scope or depth of the curriculum as well as modifying the pedagogical and educational process, depending on student achievements.

A lot of research observes academic success generally, such as success in individual courses or groups of courses or in the individual phases of studying, all in terms of current variables such as commitment to studying, fulfillment of obligations, quality of delivered educational processes, perceived difficulty of the curriculum and different socio-demographic variables (place of residence, gender, income, habits). Rarely have we undertaken scientific observation of success in high school, especially in individual subjects, or success in the state exam and completion of the high school curriculum. It is our opinion that these factors can have a big influence on students' success in the early phase of higher education because they contain acquired knowledge, work habits and attitudes towards studying. Therefore, success in high school is included in the suggested model in order to investigate its influence on the output variable.

The aim of this paper is to create a model that will successfully classify students into one of two categories, depending on their success at the end of their first academic year, as well as finding meaningful variables affecting student success. The methodology used is a decision tree method. Several classification algorithms were used in the process. The decision trees method, as well as data used for building a model, are described in the chapter Methodology.

This paper presents an overview of previous research in this particular field, the methodology used for assembling the model, research results and a conclusion with some guidelines for future research.

2. Previous research

In the research of Zekić-Sušac et al. [29], models for predicting student success were devised using decision trees and neural networks. Research was conducted among students from the second, third and fourth academic years. The sample consisted of 165 students. The output variable of the model was Grade average from the previous academic year expressed as two categories – less than or equal to 3 and greater than 3, whereas the input variables were Gender, Scholarship,

Time dedicated to studying, Exam materials, Students taking preliminary exams, Lecture attendance by students, Attendance of exercises by students, Importance of the achieved exam grade. The average classification rate with a decision tree was 88.36%, and analysis of the significance of input variables indicated that the variable Time dedicated to studying was the most significant. Neural networks gave a lower rate of accuracy (66.26%), and the statistical t-test showed a statistically significant difference.

A paper written by Jadrić et al. [10] explored the use data mining methods in higher education and created classification models for neural networks, decision trees and logistic regression. The analysis was carried out on a sample of 715 students. The results indicate that women drop out less than men, students who have attended high schools drop out less often than those who have attended other schools, and generally speaking, students with better entrance rankings drop out less. The neural network model was evaluated as the best when compared to all the other models.

Vandamme et al. [26] conducted research aimed at predicting academic performance of first-year students. The aim of the paper was to classify students into three groups: (1) low-risk students exhibiting a high probability of success, (2) medium-risk students who may succeed if the university takes appropriate measures, and (3) high-risk students who have a high probability of failing or dropping out. The research endeavored to classify the students into these three groups, prior to sitting for the first-year exams, which would have made it easier to assist them. The research sample comprised 533 students. Student classification algorithm ID3 was chosen with 5 input variables. The correct classification rate was 40.63%, specifically, 48.65% for high-risk students, 18.46% for medium-risk students and 60.34% for low-risk students. The most significant variables were Weekly course attendance by students and their Feeling of having made a good decision to enroll into the particular university.

Using decision trees and neural networks, Cheewaprakobkit [3] constructed a model in to classify students based on their academic achievement. The dataset comprised 1,600 student records with 22 attributes of students enrolled between 2001 and 2011 at a university in Thailand. A cross-validation with 10 folds was used to evaluate the prediction accuracy. The results showed that the decision tree classifier achieves a high accuracy of 85.188%, which is 1.313% higher than the neural network classifier.

The aim of Shah's [24] study was to investigate factors affecting the academic performance of students by comparing the accuracy of different classifiers. Students were categorized in five groups based on performance such as: “very good” students - a high probability of succeeding; “good” students - above average results with little effort and who may succeed with good grades; “satisfactory” students – those who may succeed; “below satisfactory” students – those invest more efforts to succeed; and “fail” students - a high probability of dropping out.

The dataset comprised 231 students. Several machine learning algorithms were used: J48, RandomForest, RepTree and BFTree of Decision Trees, Bayes and NaiveBayes of BayesNetworks, Logistic and RBFNetwork functions and the JRip rule. The best result was given by the BayesNet algorithm with an accuracy of 51%. Data was then resampled using the Weka resample function. This function oversampled the minority class and undersampled the majority class. The resampled dataset was significantly more accurate. RandomForest turned out to be the most effective classifier (with a 92% accuracy).

Ibrahim and Rusli [9] compared an artificial neural network, decision tree and linear regression to predict the academic performance of students. The academic performance indicator in this study was measured using the cumulative grade point average (CGPA) at graduation. The demographic profile of students and the CGPA for the first semester of undergraduate studies were used as the predictor variable for the academic performance of undergraduate students. The result of this study showed that all three models had an accuracy of more than 80%.

Osmanbegović and Suljić [18] compared various data mining methods and techniques in predicting student success, applying survey data collected from first-year students and enrolment data. The sample included 257 students. Student success was based on their grades in the Business Informatics course, which was also the output variable. Input to the model comprised 12 variables (gender, high school, scholarships, materials, grade importance, earnings, etc.). The Naive Bayes had a better prediction than other algorithms (i.e. J48 and Multilayer Perceptron) with an accuracy of 76.65%.

The aim of Simeunović and Preradović [25] in their paper was to devise a model for predicting the student performance using data mining. The model created using the student socio-demographic data, behavioral data, personality characteristics, attitudes towards learning and the entire teaching process organization, tends to classify students into one of two categories of success. Performance was measured using the student grade point average achieved over the course of studies. They tested three data mining methods: logistic regression, decision trees and neural networks. The decision trees exhibited an accuracy of 71.25%, logistic regression 74.8%, and neural networks 76.4%.

Another type of research was conducted by Herzog [7], who, with the help of decision trees and neural networks, endeavored to estimate student retention and degree-completion time. According to this author, the ability to identify those students at risk of dropping out or who is are to take an exceedingly long time to graduate facilitates to direct intervention programs to where they are needed most and offers ways of improving enrollment, graduation rates, and precision of forecasting tuition revenue. His study compares the prediction accuracy of three decision trees and three artificial neural networks with that of multinomial logistic regression. Retention predictions were based on the second-year enrollment of

8,018 new full-time freshmen, and dataset used for the "time to degree" (TTD) completion comprised 15,457 records. Forty predictors were used to estimate retention, and seventy-nine variables were included in the more complex TTD forecasts. After excluding student transfers and using the C5.0 algorithm, the author achieved an accuracy of 93%.

Nghe et al. [16] compared two data mining techniques, the decision tree and the Bayesian network. They used student records and grade point average at the end of second year to predict performance in third year. Decision trees had an accuracy of 94.03% and Bayesian network an accuracy of 90.27%.

Kovačić [12] tried to predict student success by mining enrolment data. The dataset covered over 450 students who enrolled into the Information System course. He used decision trees and logistic regression. Among the decision tree growing methods, CART was most successful with an overall 60.5% percent of correct classifications.

Yadav and Pal [28] applied decision tree algorithms to data on engineering students to predict their performance in the final exam. The dataset comprised 90 student records with 16 attributes. The C4.5 algorithm produced the best accuracy standing at 67.78%.

Delen [4] developed an analytical model to predict and explain the reasons behind the attrition of freshmen students, using five years of institutional data including several data mining techniques, such as neural networks, decision trees, support vector machines and logistic regression. The sensitivity analysis of the models revealed that the educational and financial variables were among the most important predictors. Based on hold-out sample results, support vector machines generated the best overall prediction with an accuracy of 81.18%, followed by decision trees, neural networks and logistic regression.

Based on the mentioned studies, it becomes evident that numerous authors used data mining techniques to predict student success. In most cases, decision trees had the highest rate of classifying successful students. In almost all studies, output variable was grade averages, expressed in two or more categories. Also, C4.5 (J48) and CART algorithms were mostly used decision tree algorithm in previous research. Therefore, this study uses decision trees, by compareing the success of several algorithms available in the Weka data mining tool.

Author/s	Year	Sample	Methodology	No of input variables	Out-put variable	Results	The most signify-cant var.
Herzog	2006	8,018 students	decision trees, neural networks	40	Student retention (two groups)	Classification rate **Decision tree: 93%** Neural network: 85%	not conducted
Vandamme et al.	2007	533 students	discriminant analysis, neural networks, decision trees	25	Average mark (three groups)	Classification rate Decision tree: 40.63% Neural network: 51.88% **Discriminant analysis: 57.35%**	Weekly attendance of courses by students, Feeling of having made a good decision to enroll into the particular universiy
Ibrahim and Rusli	2007	206 students	neural networks, decision trees, linear regression	4	Cumulative Grade Point Average	Square root of average squared error Decision tree: 0.1769 **Neural network: 0.1714** Linear regression: 0.1848	not conducted
Nghe et al.	2007	20,492 students	decision trees, Bayesian network	14	Grade Point Average (two groups)	Classification rate **Decision tree: 94.03%** Neural network: 90.27%	Cumulative Grade Point Average Year2, English Skill
Zekić Sušac et al.	2009	165 students	decision trees, neural networks	8	Grade average (two groups)	Average classification rate **Decision tree: 88.36%** Neural network: 66.26%	Time dedicated to studying

Jadrić et al.	2010	715 students	neural networks, decision trees, logistic regression	24	Dropout (two groups)	-	not conducted
Delen	2010	7,018 students	neural networks, decision trees, support vector machines, logistic regression	39	Second Fall Registered (two groups)	Classification rate Decision trees: 80.65% Neural network: 79.85% **SVM: 81.18%** Logistic regression: 74.26%	Student attendance, Student loans
Shah	2012	231 students	decision trees, Bayesian networks, functions (logistic, RBFNetwork), JRip rule	5	Grade Point Average (five groups)	Classification rate **Decision tree: 92%** Bayesian network: 59% Logistic function: 66% JRip rule: 75%	Academic integration, family background, social integration
Osmanbegović and Suljić	2012	257 students	neural networks, decision trees, Bayes network	12	Success in the course "Business informatics" (two groups)	Classification rate **Bayes network: 76.65%** Neural network: 71.20% Decision tree: 73.93%	not conducted
Kovačić	2012	453 students	decision trees, logistic regression	9	Study outcome (two groups)	Classification rate **Decision tree: 60.5%** Neural network: 59.4%	not conducted

Yadav and Pal	2012	90 students	decision trees	16	Student result in first year of Engineering (three groups)	Classification rate **Decision trees: 67.78%**	not conducted
Simeunović and Preradović	2014	354 students	decision trees, logistic regression, neural networks	17	Grade Point Average (two groups)	Classification rate Decision tree: 71.25% Logistic regression: 74.8% **Neural network: 76.4%**	not conducted
Cheewaprakobkit	2015	1,600 students	decision trees, neural networks	20	Cumulative Grade Point Average (five groups)	Classification rate **Decision trees: 85.19%** Neural network: 83,88%	The number of hours worked per semester

Table 1: *An overview of previous research*

3. Methodology

3.1. Decision tree

The decision tree is a data mining technique for solving classification and prediction problems. Decision trees are a simple recursive structure for expressing a sequential classification process in which a case, described by a set of attributes, is assigned to one of a disjoint set of classes. Decision trees consist of nodes and leaves. Each node in the tree involves testing a particular attribute and each leaf of the tree denotes a class. Usually, the test compares an attribute value with a constant. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached, the instance is classified according to the class assigned to the leaf. If the attribute that is tested at a node is a nominal one, the number of children is usually the number of possible values of the attribute. The tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used [15] [27] [20] [23].

As mentioned before, the problem of constructing a decision tree can be expressed recursively. First, it is necessary to select an attribute to place at the root node, and make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute. Now the process can be repeated recursively for each branch, using only those instances that actually reach the branch. If at any time all instances at a node have the same classification, that part of the tree has to stop developing [27]. According to Vandamme [26], the way finding the attribute that produces the best split in the data is the one of the main differences between the various decision-tree-building algorithms.
There are several measures of splitting criteria. Each decision tree algorithm use its own measure to select among the attributes at each step while growing the tree.
Several decision trees algorithms were used in this research, as it is described later in the paper.

ID3 was designed for cases where there are many attributes and the training set contains many objects, but where a reasonably good decision tree is required without much computation. It has generally been found to construct simple decision trees, but the approach it uses cannot guarantee that better trees have not been overlooked [22]. ID3 learns decision trees by constructing them top-down. Each instance attribute is evaluated using a statistical test to determine how well it alone classifies the training examples. The best attribute is selected and used as the test at the root node of the tree. The entire process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree. This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices [15].
According to Mitchell [15], the central choice in the ID3 algorithm is selecting which attribute to test at each node in the tree. There is a good quantitative measure for this problem, called information gain. But in order to define information gain precisely, it is necessary to define a measure commonly used in information theory, called entropy, that characterizes the (im)purity of an arbitrary collection of examples.
If the target attribute can take on m different values, then the entropy of S relative to this m-wise classification is defined as [15]:

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2 p_i \tag{1}$$

Where S is a given collection and p_i is the proportion of S belonging to class i.

The given entropy as a measure of the impurity in a collection of training examples, a measure of the effectiveness of an attribute in classifying the training data can be defined now. The measure is called information gain. It is the expected reduction in entropy caused by partitioning the examples according to this attribute. The information gain, *Gain(S,A)* of an attribute A, relative to a collection of examples S, is defined as:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

where *Values(A)* is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v.

The **C4.5** algorithm was proposed in 1992, by Ross Quinlan, to overcome the limitation of the ID3 algorithm (unavailable values, continuous attribute value ranges, pruning of decision trees, etc.) [8]. C4.5 uses a divide-and-conquer approach to growing decision trees. The default splitting criterion used by C4.5 is gain ratio, an information-based measure that takes into account different number of test outcomes [21].

$$GainRatio(S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \quad (3)$$

The **J4.8** algorithm is Weka's implementation of the C4.5 decision tree learner (J4.8 actually implements a later and slightly improved version called C4.5 revision 8, which was the last public version of this family of algorithms before the commercial implementation C5.0 was released) [27].

REPTree (Reduced Error Pruning Tree) builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as does C4.5 [27]. RepTree uses the regression tree logic and creates multiple trees in different iterations. After that, it selects the best one from all generated trees. That is then considered the representative. In pruning the tree, the measure used is the mean square error on the predictions made by the tree. REPTree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using the C4.5 method of using fractional instances [11].

RandomTree is an algorithm for constructing a tree that considers K random features at each node. It performs no pruning [27] (cited in [19]).

RandomForest is a combination of tree predictors where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [2]. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting [2] (cited in [13]).

For making decision trees, researchers used the Weka system for testing datasets using a variety of open source machine learning algorithms. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It was developed at the University of Waikato in New Zealand. The workbench includes methods for the main data mining problems: regression, classification, clustering, association rule mining, and attribute selection [27].

3.2. Data

Data for composing a model was collected from two sources during the period of three academic years (2015/16, 2014/15 and 2013/14):

- Data from the ISVU system[‡] of students enrolled in the second academic year
- State exam results[§] of those same students after graduating from high school

We took into consideration only students from the ISVU system who enrolled into the second academic year because there is a weighted grade average calculated for their first academic year. The data collection was sourced from student data (personal identification number), high school and programs they had completed, average high school grade, grades from the state exam (as a sum of grades in subjects demanded by the university faculty) and separately, grades in individual state exam subjects (Croatian language, mathematics and foreign languages).

[‡] ISVU (Information System of Higher Education Institutions) is a project of Ministry of Science, Education and Sport, launched in mid-2000 as a part of the informatization program of higher education institutions in the Republic of Croatia. The system has all relevant data about students, professors, courses, exams and information resources.

[§] State exam is a collection of exams which are conducted under equal conditions and criteria for all students at the same time and it enables obtaining comparable results of students on a national level. It is conducted by taking state exams. [14]

This dataset also comprised information on rank of importance that students gave to each faculty (1-10) in which they planned to enroll after finishing high school (detailed analysis showed that for almost all students, the Faculty of Economy in Osijek was the first or the second choice in rank of importance, hence this variable was excluded from the model).
Data refers to students from the Faculty of Economy (University of Osijek) for academic years 2014/15, 2013/14 and 2012/13.
After the process of cleaning up and eliminating useless and incomplete data, data from two different sources were merged using the common field (personal identification number) and the outcome sample comprised 665 students.

The decision tree model used 8 input variables, as shown in Table 2. Data for the first seven variables was collected from state exam results and the student enrollment status, with the output variable, weighted grade average collected from the ISVU system. The input variable program refers to the high school education program which students finished. There were several high school programs, such as those for economists, grammar schools, various types of technicians (ecological, agroturism, pharmaceutical, geodesic, building, medical, etc.), graphic designer, business secretary, commercial school, and so on. Since the proportion of these programs in the total sample is only 9.5%, they were combined into a single group, called "Other". Hence, there are three groups of variable programs – Economist, Grammar school and Other. The Variable Highschool refers to points achieved by a student based on the grade point average in high school. Points that students achieved at the state exam examination are represented by the variable StateExam. Total points from the state exam (points from high school and exams), are represented by the variable Overall. The variable ForeignLang represents percentage of success in the English or German language, the variable CroatianLang shows the percentage of success in Croatian language, while the variable Math shows percentage of success in mathematics. By taking the examination in the Croatian language, mathematics and a foreign language, students can choose the exam level (A – higher or B – basic) according to the requirements of particular faculties. If a student passed both exam levels, the higher achieved points were recorded in the dataset. The variable Status refers to student enrollment status (full-time or part-time student).
For numerical variables minimum, maximum, mean and standard deviation were calculated.

No.	Variable	Description	Frequency/statistics
1	Program	Highs school education program finished	1 – Economist (51.00%) 2 – Grammar school (39.50%)

			3 – Other (9.50%)
2	Highschool	State exam – points from high school	Min: 94.8 Max: 199.2 Mean: 158.068 StdDev: 21.565
3	StateExam	State exam – points on the exam	Min: 236.47 Max: 623.32 Mean: 437.348 StdDev: 67.524
4	Overall	State exam – total points	Min: 345.3 Max: 819.3 Mean: 595.419 StdDev: 75.059
5	ForeignLang	State exam – % of success in foreign language	Min: 35.5 Max: 94 Mean: 69.506 StdDev: 11.981
6	CroatianLang	State exam – % of success in Croatian language	Min: 43.13 Max: 97.5 Mean: 68.012 StdDev: 8.578
7	Math	State exam – % of success in mathematics	Min: 21.67 Max: 92.5 Mean: 60.067 StdDev: 17.18
8	Status	Status of students' enrollment	1 – Full-time (88.00%) 2 – Part-time (12.00%)

Table 2: *Input variables*

The weighted grade average of students after finishing the first year of study was chosen as the output variable. The variable was expressed as a nominal with two classes – BELOW (average $<$ 3.5) and ABOVE (average $>=$ 3.5) which transforms the mentioned problem into classification problem.

For the needs of composing a decision tree model, data were divided into a training sample and testing sample. A total sample of 665 students was first filtered with the variable Weighted average, according to which 159 (24%) students had an average higher or equal 3.5, and the remaining 506 (76%) had an average lower than 3.5. The equal distribution of students was taken into account in order to create a more successful model. Since the total sample consists of a larger number of students with a lower average, the training sample consists of

2/3 of students with average higher or equal 3.5 which equals about 100 units, and the same number of students with an average less than 3.5. Accordingly, the training sample equals 200 students. For the needs of testing the decision tree model, a special file was created, comprising the remaining 465 students. The structure of samples for training and testing is shown in Table 3:

Sample	ABOVE	BELOW	Total
Training	100 (50.00%)	100 (50.00%)	**200 (100.00%)**
Testing	59 (12.69%)	406 (87.31%)	**465 (100.00%)**
Total	**159 (23.91%)**	**506 (76.09%)**	**665 (100.00%)**

Table 3: *Structure and division of samples*

4. Results

As a measure of success of the model, the classification rate was used on the testing sample. For composing a decision tree model, several algorithms were used, where their functioning is described in the chapter Methodology. The results obtained using tree decision method are shown in Table 4:

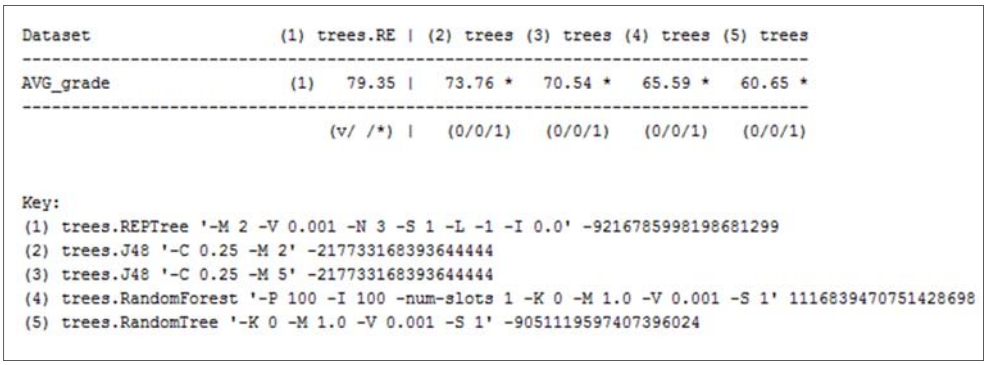
Decision tree algorithm	MinNumObj*	Number of Leaves	Size of the tree	Correctly Classified Instances	Incorrectly Classified Instances
J48	2	17	32	343 (73.76%)	122 (26.24%)
J48	5	8	14	328 (70.54%)	137 (29.46%)
RandomForest				305 (65.59%)	160 (34.41%)
RandomTree			114	282 (60.65%)	183 (39.35%)
REPTree			7	369 **(79.35%)**	96 **(20.65%)**

* The minimum number of instances per leaf

Table 4: *Decision tree results*

Table 4 shows that the REPTree algorithm had the highest classification accurate rate of 79.35%. Using the Weka Experiment Environment, a statistical test of significance for one learning scheme (REPTree) against four others was conducted. The test showed that there is a statistically significant difference

between the REPTree algorithm and all other algorithms, and that the REPTree algorithm is significantly better than others at the level of 95% (Figure 1).



```
Dataset                  (1) trees.RE | (2) trees (3) trees (4) trees (5) trees
-------------------------------------------------------------------------------
AVG_grade                 (1)   79.35 |   73.76 *   70.54 *   65.59 *   60.65 *
-------------------------------------------------------------------------------
                               (v/ /*) |   (0/0/1)   (0/0/1)   (0/0/1)   (0/0/1)


Key:
(1) trees.REPTree '-M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0' -9216785998198681299
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(3) trees.J48 '-C 0.25 -M 5' -217733168393644444
(4) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698
(5) trees.RandomTree '-K 0 -M 1.0 -V 0.001 -S 1' -9051119597407396024
```

* significantly worse

Figure 1: *T-test results on decision trees*

At first glance, the decision tree quite correctly classified students according to success after the first year of study. However, upon more accurate inspection of Table 5, what is noticeable is that the tree is especially successful in recognizing students with a lower grade average (86%), while this is not the case for students with a higher average, where the rate of accurate classification is only 32%.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
above	0.322	0.138	0.253	0.322	0.284	0.167	0.587	0.164
below	0.862	0.678	0.897	0.862	0.879	0.167	0.587	0.895

Table 5: *Detailed accuracy by class (REPTree algorithm)*

Given that there is a relatively large disproportion between the rates of accurate classification of some classes, the authors decided to calculate the average rate of classification for two models that show the highest rate of accurately classified instances. These are models that use the REPTree and J48 algorithm. With the J48 algorithm, calculations use the one with 5 instances per leaf because it gives a smaller tree. Average rates of classification are shown in Table 6:

Decision tree algorithm	Rate of classification, class "above" (%)	Rate of classification, class "below" (%)	Average rate of classification (%)
REPTree	32.2	86.2	59.2
J48	47.5	73.9	**60.7**

Table 6: *Calculated results of average rate of classification*

The highest average rate of classification in this case is given with a tree that uses J48 classification algorithm. We consider this model is better for determining the academic success regardless of the fact that the total rate of classification is lower, because it is better to have a bigger accuracy of classification for both classes for this type of observed problem.
The result on a test set is often displayed as a two-dimensional *confusion matrix* with a row and column for each class. Each matrix element shows the number of test examples for which the actual class is the row and the predicted class is the column. Good results correspond to large numbers down the main diagonal and small, ideally zero, off-diagonal elements [27].

		Predicted Class	
		above	below
Actual Class	above	28	31
	below	106	300

Table 7: *Confusion matrix*

Table 7 shows the confusion matrix on the testing sample, where it becomes clear that of the total of 59 students with average greater than or equal to 3.5, the decision tree managed to place 28 of them into the correct category. Regarding the class of students with an average less than 3.5, the decision tree managed to accurately place 300 students, whereas 106 were placed into student classes with higher averages.

The structure of the composed decision tree is shown in Figure 2.

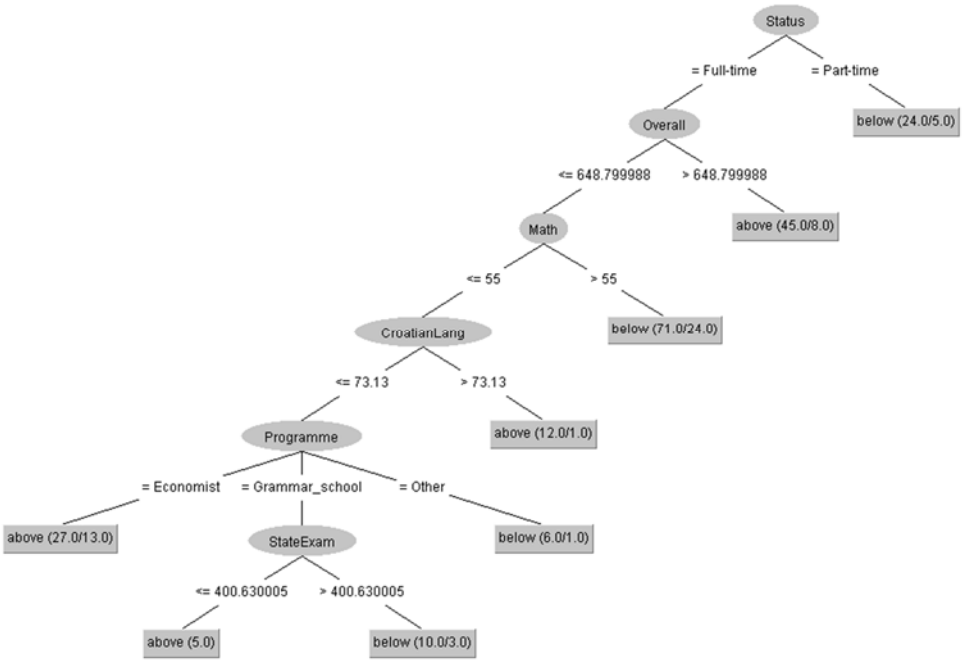


Figure 2: *The decision tree obtained using algorithm J48*

This tree consists of 6 nodes and 8 leaves and it branches to the left. The first splitting node is the variable Status. For a part-time student, the weighted average grade after finishing first year of study will be less than 3.5. Otherwise, the tree continues to split. The next splitting node is at the variable Overall, meaning that if a student achieved more than 648 (out of 1000) points on the state exam, the weighted average grade will be greater than 3.5. The next nodes are Math and CroatianLang, followed by Program and StateExam.

4.1. Attribute selection

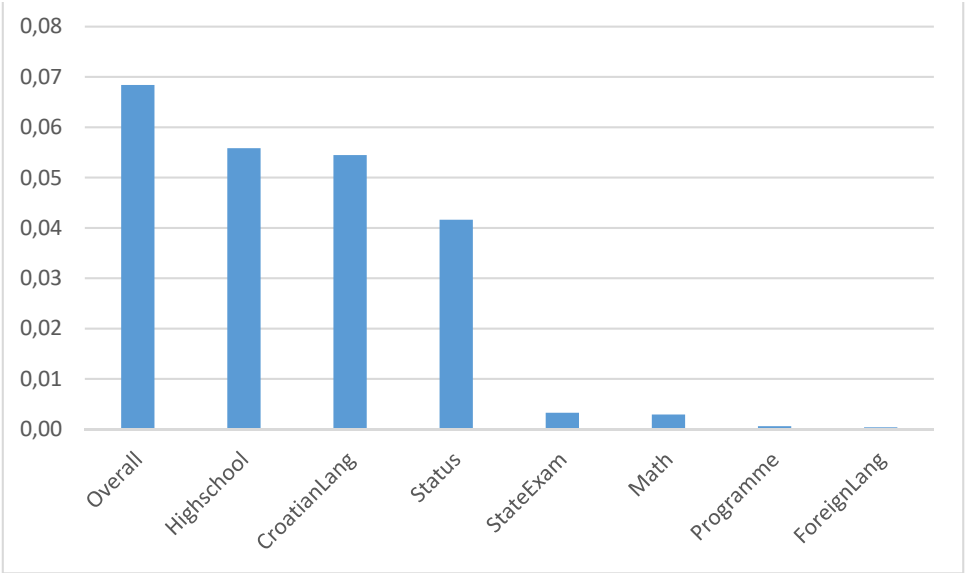
Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. To do this, two objects must be set up: an attribute evaluator and a search method. The evaluator determines what method is used to assign a worth to each subset of attributes. The search method determines what style of search is performed [1].

According to Hall and Holmes [6], referent techniques of attribute selection are information gain and Relief, while Ganchev et al. [5] considers that they are Information Gain and Gain Ratio (cited in [17]). Therefore, attribute evaluation in this paper takes 3 methods into consideration: Information Gain, Gain Ratio and Relief. Since different methods give different attribute selection results, the

average value of all methods used was taken as a final result of attribute ranking. The ranker method was used as a search method.

Attribute	Information Gain	Gain ratio	Relief	Average
Programme	0.00594	0.00441	-0.0085	0.000617
Highschool	0.06653	0.07716	**0.02382**	0.055837
StateExam	0	0	0.00984	0.00328
Overall	**0.08444**	**0.10853**	0.01224	**0.068403**
ForeignLang	0	0	0.0012	0.0004
CroatianLang	0.07659	0.0809	0.00592	0.05447
Math	0	0	0.00878	0.002927
Status	0.03543	0.06692	0.0225	0.041617

Table 8: *Results of analysis of input variables' significance*



Graph 1: *Graphical representation of the significance of input variables*

The test results are shown in Table 8, and the order of the most significant variables is shown in Graph 1. The variable Overall affects the output the most, followed by Highschool, CroatianLang and Status. The variables StateExam, Math, Programme and ForeignLang have almost no influence on the output variable. Analysis has shown that the Total points obtained at state exam, Points from high school and Points obtained at state exam in the Croatian language are the most influential variables for predicting the academic success.

5. Discussion and conclusion

Several models of decision trees for classifying the academic success of students were devised in this research using several notable classification algorithms. It has been shown that data mining tools can largely be used by education institutions for predicting student success. The highest classification rate of 79% was produced using the REPTree decision tree algorithm. Regardless of the high accuracy, the tree was not equally successful in classifying both classes, hence it gave weak results in recognizing students with an average greater than 3.5. Therefore, the average rate of classification was calculated for two models that gave out the highest total rate of classification, whereupon there is a higher percentage achieved by the model that used algorithm J48. Given that the difference between average rates of classification is very small (only 1.5%), it becomes necessary to expand the research to see whether the results will repeat.

The most significant variables were Total points on state exam that included points on the examination itself in addition to points that the candidate achieved based on the average grade in high school, points from high school and points on Croatian language exam. The total result on the state exam is the expected result, however, surprisingly, that the Success on the state exam is a variable with a low level of significance, and Average grade in high school is a variable with a high level of significance, although the Total success on the state exam was contributed by ¼ of its value. Equally so, unexpectedly, the Success on the state exam in the Croatian language was ranked third according to its significance. The expectation was also that Success in mathematics on the state exam would have a higher significance.

There is also the question dependence between variables, and if so, which have an influence on the structure of decision trees, i.e., the significance of variables. Therefore, linear dependencies between the variables Overall, StateExam, Highschool, CroatianLang, ForeignLang and Math were studied using the Pearson coefficient with $p<0.01$). A high correlation coefficient (0.962) was obtained only between the variables Overall and StateExam, while among the other variables, the coefficients were moderate or weak (0.5-0.63 and less than 0.25). Accordingly, one might expect that the StateExam variable would have a significance close to the significance of the variable Overall, which was not the case. Collinearity tests of variables cannot be linked to relationships of significance for individual variables measured based on their value of information content, which means a non-linear relationship between variables. Although the expectation is that academic success in the first year of study would depend on the high school program, this was not the case. It should be noted that decision trees are based on non-linear methods, hence possible correlations between variables have not effect on the decision tree model.

It is difficult to say whether students who achieved a certain (better or worse) average at the end of the first academic year will continue with the same success in the following years.
The limitation of this study is that students may enroll into second year without passing all first-year courses. In other words, they may sit for the exams in the following year. Therefore, the weighted grade average of these students can be changed (probably not significantly).
When used in various research papers by other authors, neural networks turned out to be a good predictor of student success. Hence, our future research will also compare this model with other particular data mining techniques, such as cluster detection, memory-based reasoning and support vector machines. The potential benefits of this model are great; however, this model needs to be improved in order to achieve a higher (total and average) accuracy rate. The model accuracy rate could be improved by introducing additional variables, such as course grades on the first year of study, conducting primary research among the student population, or by even increasing the sample size.

References

[1] Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. (2016). WEKA Manual for Version 3-8-0.

[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[3] Cheewaprakobkit, P. (2015). Predicting student academic achievement by using the decision tree and neural network techniques. Catalyst, 12(2), 34-43.

[4] Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

[5] Ganchev, T., Zervas, P., Fakotakis, N. and Kokkinakis, G. (2006). Benchmarking Feature Selection Techniques on the Speaker Verification Task. Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing, 314-318.

[6] Hall, M. A. and Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering, 15(3), 1437-1447.

[7] Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. New Directions for Institutional Research, 2006(131), 17-33.

[8] Hssina, B., Merbouha, A., Ezzikouri, H. and Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications, 4(2), 13-19.

[9] Ibrahim, Z. and Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. 21st Annual SAS Malaysia Forum, 5th September 2007, Kuala Lumpur.

[10] Jadrić, M., Garača, Ž. and Ćukušić, M. (2010). Student dropout analysis with application of data mining methods. Management, 15(1), 31-46.

[11] Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. IJISET - International Journal of Innovative Science, Engineering & Technology, 2(2), 438-446.

[12] Kovačić, Z. (2012). Predicting student success by mining enrolment data. Research in Higher Education Journal, 15(1).

[13] Liaw, A. and Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

[14] Ministry of Science, Education and Sport (2009). The State Exam.

[15] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[16] Nghe, N. T., Janecek, P. and Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *37th ASEE/IEEE Frontiers in Education Conference*, 10-13th October 2007, Milwaukee.

[17] Oreški, D. (2014). Evaluation of contrast mining techniques for feature selection in classification. doctoral thesis, Varaždin: Faculty of Organization and Informatics.

[18] Osmanbegović, E. and Suljić, M. (2012). Data mining approach for predicting student performance. Economic Review: Journal of Economics & Business / Ekonomska Revija: Casopis za Ekonomiju i Biznis, 10(1), 3-12.

[19] Ozer, P. (2008). Data Mining Algorithms for Classification. BSc Thesis Artificial Intelligence, Radboud University Nijmegen.

[20] Quinlan, R. J. (1987). Generating Production Rules from Decision Trees. *IJCAI,* 87, 304-307.

[21] Quinlan, R. J. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Arti, 4, 77-90.

[22] Quinlan, R. J. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[23] Rokach, L. and Maimon, O. (2014). Data Mining with Decision Trees: Theory and Applications. World scientific.

[24] Shah, N. S. (2012). Predicting Factors That Affect Students' Academic Performance by Using Data Mining Techniques. Pakistan Business Review, 13(4), 631-638.

[25] Simeunović, V. and Preradović, Lj. (2014). Using data mining to predict success in studying. Croatian Journal of Education, 16(2), 491-523.

[26] Vandamme, J.-P., Meskens, N. and Superby, J.-F. (2007). Predicting Academic Performance by Data Mining Methods. Education Economics, 15(4), 405-419.

[27] Witten, I. H., Frank, E. and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann Publishers.

[28] Yadav, S. K. and Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT), 2(2), 51-56.

[29] Zekić-Sušac, M., Frajman-Jakšić, A. and Drvenkar, N. (2009). Neuron networks and trees of decision-making for prediction of efficiency in studies, Ekonomski vjesnik (Econviews), 22(2), 314-327.